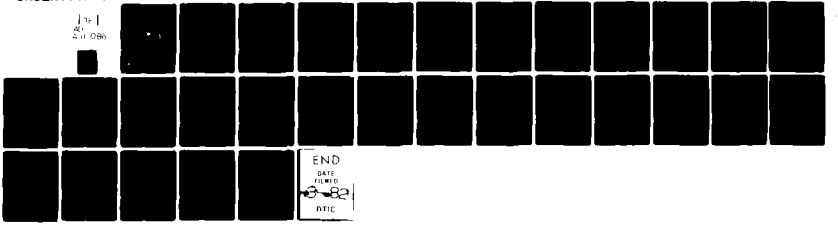
AD-A111 086 NAVAL BIODYNAMICS LAB NEW ORLEANS LA F/G 5/10
STATISTICAL ISSUES IN PERFORMANCE TESTING: COLLECTED PAPERS.(U)
SEP 81 A C BITTNER, W B JONES, R C CARTER
UNCLASSIFIED NBDL-81R010 NL



NBDL - 81R010

LEVEL II

30

STATISTICAL ISSUES IN PERFORMANCE TESTING: COLLECTED PAPERS

Alvah C. Bittner, Jr., Marshall B. Jones, Robert C. Carter, Richard H. Shannon, Douglas C. Chatfield, Robert S. Kennedy

September 1981

NAVAL BIODYNAMICS LABORATORY
New Orleans, Louisiana

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NBDL-81R010	2. GOVT ACCESSION NO. AD-A111 086	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Statistical Issues in Performance Testing: Collected Papers		5. TYPE OF REPORT & PERIOD COVERED Research Report 6. PERFORMING ORG. REPORT NUMBER NBDL-81R010
7. AUTHOR(s) A. C. Bittner, Jr., M. B. Jones, R. C. Carter, R. H. Shannon, D. C. Chatfield, R. S. Kennedy		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Biodynamics Laboratory PO Box 29407 New Orleans, LA 70189		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Project F58524 Task Area ZF5852406 Work Unit MF58.524-002-5027
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Medical Research and Development Command Bethesda, MD 20014		12. REPORT DATE September 1981 13. NUMBER OF PAGES 26
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release, distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Repeated Measures, Human Performance Testing

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This is a collection of papers about various statistical issues that arise in human performance testing. Two papers by M. B. Jones discuss practice-associated changes of individual differences in test performance. A. C. Bittner, Jr., and R. H. Shannon each present ideas about how to assess whether there is any practice-associated change of individual differences. R. C. Carter proposes the use of time series analysis to study performance of an individual. A paper by A. C. Bittner, Jr. and D. C. Chatfield shows how signal detection theory can be applied to performance testing.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

NBDL - 81R010

STATISTICAL ISSUES IN PERFORMANCE TESTING: COLLECTED PAPERS

Alvah C. Bittner, Jr., Marshall B. Jones, Robert C. Carter, Richard H. Shannon, Douglas C. Chatfield, Robert S. Kennedy

September 1981

Bureau of Medicine and Surgery
Work Unit No. MF58.524-002-5027

Approved by

Channing L. Ewing, M. D.
Chief Scientist

Released by

Captain J. E. Wenger MC USN
Commanding Officer

Naval Biodynamics Laboratory
Box 29407
New Orleans, LA 70189

Summary Page

PROBLEM

Human performance testing results in scores which represent the performance. The scores indicate differences among people, alterations due to types of stimuli within a test (e.g., changing signal intensity), effects of changes in the test environment, and, if the tests are repeated, effects of practice. Mathematical descriptions of these differences and changes are compared with the data to indicate which types of effects occurred. The mathematical models are usually statistical, due to the variability of the effects. The problem is that the utility of the test scores is limited by the generality and accuracy of the statistical-mathematical models used to interpret the data.

FINDINGS

1. Correlations between tests at each stage of practice can be useful to show changes of what is measured by the tests.
2. There are many ways to detect changes of individual differences during practice (e.g., Chi-square statistics, graphical methods, factor analysis, analysis of variance). None of the techniques studied is entirely satisfactory.
3. Signal detection theory can be useful for analysis of performance tests involving comparisons of stimuli with a standard stimulus.
4. Time series analysis can be used to explain how performance changes over time.

RECOMMENDATIONS

Human performance data should be compared with some sort of model or hypothesis about effects represented by the data. Several useful models are presented, and their application is recommended in appropriate contexts.

Trade names of materials or products of commercial or non-goverment organizations are cited where essential for precision in describing research procedures or evaluation of results. Their use does not constitute official endorsement or approval of the use of such commercial hardware or software.

This research work was funded by the Naval Medical Research and Development Command and by the Biological Sciences Division of the Office of Naval Research.

Table of Contents

Each of these papers was presented at a professional meeting or symposium. Acknowledgement of previous publication appears at the beginning of each paper.

Proceedings of the Seventh Psychology in the DOD Symposium
USAF Academy, Colorado Springs, CO
17 April 1980

Video Games and Convergence or Divergence with Practice

Marshall B. Jones
The Pennsylvania State University
Hershey, PA 17033

CDR Robert S. Kennedy, MSC, USN

and

Alvah C. Bittner, Jr.
Naval Aerospace Medical Research Laboratory Detachment
New Orleans, LA 70189

Abstract

Video Games

In 1972 a coin-operated video game called Pong and manufactured by Atari, Inc., a company founded that same year, appeared on the electronic-games market. In less than a year Atari sold 6,000 games at more than $1,000 apiece. Midway Manufacturing Co., which Atari licensed to produce a version of Pong, sold 9,000 of the table-tennis type games in less than six months.

Also in 1972, Magnavox marketed a video game called Odyssey that could be played on home TV sets. The Odyssey set included a control unit, which attached to a home TV set and permitted one to play 12 different games by inserting a "game card" into the control unit. The original Odyssey was not, however, a programmable video game. All 12 games were resident in the control unit and were not, in fact, very different; the "game card" set appropriate lines, bars, and cursors. Then in 1975 Atari entered the home video market with a version of Pong that offered several new advances: electronically generated on-screen courts, sound effects for every hit, miss, and ricochet, and automatic on-screen digital scoring. By the end of 1976 twenty different companies, including Coleco, First Dimension, National Semiconductor, Phoenix, Unisonic, and Universal Research were producing video games for home use.

About this time, that is, late in 1976, Fairchild Camera and Instrument entered the field with the first fully programmable video system. The system was programmed by inserting an electronic cartridge into the game console. The benefit was that one could play as many different games as the company provided cartridges -- in fact, more, because most cartridges contained several games. Different games within the same cartridge were selected by punching in a number on the control console.

In 1977 and 1978 programmable video games for home use proliferated on all sides. Companies already in the field, like Atari and Magnavox, came out with programmable video systems; and new companies entered the field, for example, RCA and Bally, the pinball-machine company. In 1978 American shoppers spent more than 200 million dollars on programmable home video games and everything pointed toward an even larger market in the future.

Video games as psychological tests

The potential of programmable video games for psychological testing is large. First, the new games involve skills and lots of them. Video games are tasks and playing them repeatedly constitutes so many trials of practice. The more a person plays the better he or she becomes, especially in the beginning; after extended practice, the gains from playing yet another game are small or non-existent. Most of the games, moreover, have a high ceiling, so high that few people come close to reaching it. Second, the new games are wonderfully self-motivating. A case can be made that for research purposes solid motivation is not all to the good. Insufficient motivation, boredom, or wavering attention may be precisely what the investigator wishes to study; and in such a case video games would not be the tasks of choice. More often, however, we are interested in skill acquisition, learning or forgetting, as distinct from performance; and where we are, insufficient or wavering motivation is quite simply a source of error. Third and last, most video games are highly speeded. In fact, this feature of the games may account for much of their appeal. In considerable measure the games are enjoyable because they operate at more or less the same speeds as we do, that is, as our brains do. Their being so fast, however, may permit them to tap aspects of human functioning that escaped us as long as we were dealing with essentially mechanical tasks (pursuit rotor, two-hand or complex coordination).

Programmable video games are equally attractive at a pragmatic level, especially for performance testing. Literally dozens of games and, in principle, hundreds or even thousands can be played with identically the same equipment; one need only insert another cartridge. Television sets are light, easily transported, and occupy little space. Furthermore, if they break down, they are easily replaced. The game console and associated cartridges are robust. The only parts of game equipment that show any appreciable tendency to break down are the joysticks, wheels, knobs, etc. that the subject manipulates; but these too are easily replaced.

Stabilization and task definition

Despite these many advantages psychologists have not rushed to study the new games or use them in prediction and performance testing. The first studies of programmable video games from a psychological standpoint were begun in the late summer of 1978 at the Navy Aerospace Medical Research Laboratory (NAMRL) in New Orleans. The purpose of the NAMRL studies was to find out whether or not the video games were suitable for inclusion in a performance test battery for environmental research.

A prime requirement of any performance test is that it stabilize. In a good performance test there comes a point in practice after which individual performance does not change in the absence of external changes. In group terms the mean follows a flat course, the variance among subjects remains the same from one trial to the next, and all correlations among stabilized trials are equal except for sampling variations. If a test satisfies these requirements, it may be used to study the impact of environmental variations on performance. If it does not, it is at best difficult to determine whether an observed change in performance is a practice effect or the result of environmental changes. An additional

requirement is that task definition (the average correlation among stabilized trials) be high, preferably greater than .90.

In the New Orleans laboratory a large number of conventional tests and, after September, 1978, video games have been studied over extended periods of practice, 15 consecutive working days, with a view to finding out how quickly, if at all, they stabilize and how well defined they are. So far nine video games have been studied in small samples (roughly 13 subjects) and one game, Air Combat Maneuvering (ACM), has been studied in roughly twice that number of subjects. All ten video games are made by Atari.

ACM is a remarkable task. The mean follows a classical learning curve, rising rapidly in the early trials and then gradually flattening out. The variance among subjects stabilizes after day 8 and the intertrial correlation after day 6. Task definition is very high, .93. In the first six days of practice, that is, prior to stabilization, the intertrial correlations show an exceptionally regular superdiagonal form. Altogether ACM not only meets the requirements laid down for it as a performance test but does so more fully than any conventional test, with one exception, studied at NAMRL. The exception is Arithmetic, a conventional test that seems to be stable from the outset; the reason, in all probability, is that arithmetical skills have been so thoroughly practiced in school and everyday life that the subjects come to the laboratory at or near asymptotic levels.

Data concerning other video games studied at NAMRL are more preliminary. It does seem, however, that some other games are as promising for performance testing as ACM. Breakout, for example, seems also to stabilize after six days, though with poor task definition, .77. It also seems that video games do not all depend on the same underlying skills and abilities since the correlations between tasks are in some cases quite low.

Convergence-divergence relations

The present report focuses on convergence-divergence relations among video games. When a task is practiced, its correlation with an external measure may increase, decrease, or remain the same, to take linear possibilities only into account. If the correlation increases, the task is said to converge on the external measure; if it decreases, the task diverges from the external measure.

Table 1 presents the cross-correlations between ACM and Breakout in 13 Navy enlisted volunteers. Each subject played 10 games of ACM a day for 15 consecutive working days, followed by 10 games of Breakout a day for another 15 consecutive working days. His score each day was the average of the 10 games played.

Now consider the row averages. These figures represent the correlation between each of the 15 days on ACM with the 15 days on Breakout considered as a whole. Testing for linear trend in a two-way analysis of variance, using the interaction between rows and columns as the error term, shows a small but significant tendency ($p < .01$) for ACM to converge on Breakout. The regression line rises by .07 from day 1 to day 15. Breakout, on the other hand, converges strongly on ACM. The regression line for the column averages rises by .33 from day 1 to day 15.

Two points are worth underscoring. First, convergence-divergence relations are not symmetrical. Because task A converges on task B it does not follow that task B converges on task A; task B may, in fact, diverge from task A. Second, Breakout followed ACM in time. Therefore, the correlations between Breakout and ACM increased with increasing temporal separation. Day 1 on Breakout followed ACM directly while day 15 came almost three weeks later. Nevertheless, the correlations with ACM increased systematically over this interval. This result is without precedent in the literature of differential psychology; in all other studies the correlation between the same or similar measures either decreases with increasing temporal separation or remains the same.

ACM and Breakout were the first two in a series of five video tasks; the other tasks were, in order, Surround, Race Car, and Slalom. The same 13 subjects practiced all five tasks. Breakout converges strongly not only on ACM but on the other three tasks as well; linear change over the 15 days is roughly the same in all four cases, on the order of .30. ACM, however, shows no change with Surround, a slight but significant divergence from Race Car and a stronger divergence from Slalom. The linear decrease from day 1 to day 15 is .06 for Race Car and .13 for Slalom. The last two cases are the obverse of the relations between ACM and Breakout. ACM precedes Race Car and Slalom. Therefore, since it diverges from these two tasks, the correlation between ACM and Race Car or Slalom decreases as ACM gets closer and closer temporally and sequentially to the two following tasks. These results are also without precedent in the differential literature.

Application to pilot selection

A test converges on or diverges from a training criterion according as the correlation between test and criterion increases or decreases with practice on the test. If the test diverges, there is plainly no point in extending practice on the test since the effect is to lower predictive validity; if it converges, however, there may be no predictive validity at all without extended practice.

Pilot training takes place in a series of stages, each one (except the first) building on at least some of the preceding stages. It is possible, therefore, to speak not only of a test converging on or diverging from the criterion but also of the criterion converging on or diverging from a test. If the correlation between flight grades, for example, and a test increases with level of training, the criterion converges on the test. If the correlation decreases as students progress to more and more advanced stages, the criterion diverges from the test. In the first case, where training criteria converge on a test, we have reason to believe that the test will predict operational performance at least as well as it does performance in training. If the training criterion diverges from a test, however, the test may easily be valid in training but much less so or not at all in operations.

TABLE 1

Cross-correlations between Air Combat Maneuvering (ACM) and Breakout in 13 Navy volunteers

	Breakout															
ACM	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	$\bar{\pi}$
1	29*	34	33	58	38	47	68	73	62	62	79	65	72	62	83	58
2	12	42	54	56	43	57	82	77	64	83	71	56	85	78	83	63
3	00	26	43	47	36	49	76	65	51	73	70	47	76	64	79	53
4	26	36	59	53	42	52	65	68	60	61	64	50	75	68	70	57
5	28	43	58	64	59	66	78	66	61	73	71	65	73	78	82	64
6	34	45	50	62	51	61	76	79	59	65	68	65	70	67	81	62
7	28	37	54	54	47	57	70	64	53	66	63	57	69	72	73	58
8	29	40	50	61	48	56	78	74	61	67	71	64	78	75	81	62
9	43	53	59	63	56	72	68	81	66	67	73	69	66	61	79	65
10	28	40	55	54	52	63	73	66	51	62	63	60	68	64	73	58
11	25	34	51	48	41	53	70	65	47	61	63	54	69	61	72	54
12	34	46	53	56	54	67	73	73	60	66	67	61	66	67	73	61
13	40	53	72	63	62	73	74	69	62	68	68	65	64	72	73	65
14	47	52	61	73	61	65	75	74	64	64	71	72	70	75	82	67
15	40	51	63	66	59	70	78	74	65	71	77	72	72	72	82	67
$\bar{\pi}$	30	42	54	59	50	61	74	71	59	67	69	61	72	69	78	61

*Decimal points omitted.

PROCEEDINGS of the HUMAN FACTORS SOCIETY - 24th ANNUAL MEETING 1980

CONVERGENCE-DIVERGENCE WITH EXTENDED PRACTICE: THREE APPLICATIONS

Marshall B. Jones

The Pennsylvania State University College of Medicine
Hershey, Pennsylvania

ABSTRACT

When a task is practiced, its correlation with an external measure may increase, decrease, or remain the same, to take only linear possibilities into account. If the correlation increases, the task is said to converge on the external measure; if it decreases, the task diverges from the external measure. This simple notion has many applications, some of them entailing important theoretical consequences. The present paper discusses three of these applications.

INTRODUCTION

The first author to recognize that practicing a task might alter its correlations with other measures was Herbert Woodrow (1939). His main finding was that the correlation between tasks tended to weaken with practice on one of them. For 15 years Woodrow's studies along this line were not pursued by other workers. Then, in the early 1950's, a series of investigations under Air Force auspices (Adams, 1953; Fleishman and Hempel, 1953; Reynolds, 1952) showed beyond any serious question that the correlations of many tasks with other measures change with practice. In general, Woodrow's early generalization held up, that is, most tasks were more strongly correlated with external measures early in practice than they were later on; but there were many exceptions. Depending on the particular task that was practiced and the particular external measure, the correlation between the two might increase, decrease, or remain the same as the task was practiced.

Recently, Jones, Kennedy, and Bittner (1980) introduced the phrase "convergence with practice" to indicate increasing correlations between a task and an external measure; similarly, "divergence" means that the correlation between a task and an external measure decreases with practice. The present paper develops this idea in three different settings: differential retention over long periods of no practice, personnel selection and classification, and the identification of latent factors.

DIFFERENTIAL RETENTION

Under constant conditions, most individuals reach a point on most tasks where they are no longer improving with practice or improving at a slow and regular rate; at this point each individual is at or near his or her asymptotic level. An array of such levels is called a terminal process (Jones, 1970a&b). At earlier points in practice an individual's level of performance can be analyzed into two parts, one reflecting the terminal process and the other individual differences in approach to terminal levels. Jones (1970a&b) calls this second part the rate process.

Suppose now that practice stops with all subjects at or near terminal levels and is followed by a long period (several months at a minimum) of no practice. When practice is resumed, many individuals will no longer be performing at terminal levels. As retraining proceeds, however, the subjects should, according to Jones' two-process theory of individual differences in skill acquisition, return to their original terminal levels. In consequence, the correlation between terminal level in original practice and performance in retraining should increase with retraining session number. Put differently, the retraining sessions should converge on terminal levels in original learning.

This consequence, it should be pointed out, has no precedent in correlations among temporally ordered measures of the same sort. The well-nigh universal rule in matrices of this description is that correlation decreases with temporal separation (Jones, 1969, 1972). Our consequence, however, calls for original learning to correlate most strongly with the temporally most removed measure, that is, the last retraining session, and least strongly with the measure closest to it in time, that is, the first retraining session.

A study to test this reasoning is currently underway at the Navy Biodynamics Laboratory in New Orleans. The design calls for 24 subjects, six tasks (all video games manufactured by Atari, Inc.), and three retention intervals (4-6 months, 10-12 months, and 16-18 months). We will consider one of these tasks, Air Combat Maneuvering (ACM), in some detail. All 24 subjects practiced ACM ten games a day for 15 consecutive working days and the ten games a subject played each day were averaged to obtain a single data point for each individual on each day; the retention interval is 16-18 months.

Air Combat Maneuvering is a remarkable task (Jones, 1979). The mean follows a classical learning curve, rising rapidly in the early trials and then gradually flattening out; the variance among subjects stabilizes after day 8. The 36 correlations among days 7 through 15 are high, $\bar{r} = .93$, and differ from one another no more than one would expect from sampling con-

siderations (Lawley's chi-squared test). In short, from day 7 on the subjects are all at or near their terminal levels, except for small amounts of random error. The average, therefore, of a subject's score on these nine days is a close estimate of that individual's terminal level.

A half dozen of these subjects have been returned to practice for five consecutive working days after 16-18 months of no practice. The question at issue is the correlation between terminal level as estimated from the average of days 7 through 15 and the five days of retraining. If Jones' theory is correct, this correlation will be lowest on day 1 of retraining and highest on day 5.

PERSONNEL SELECTION

The second application concerns personnel selection in cases where "the criterion" develops in a series of stages or phases. Pilot training, for example, takes place in stages (pre-solo, precision, acrobatics, etc.), and each student who completes training receives a flight grade for each stage. In such a case we have two kinds of convergence-divergence to consider. First, does a predictor task converge on or diverge from the flight training criterion, taken, let us say, as the average flight grade in advanced training? If it converges on the criterion, then predictive validity increases with practice. If the task diverges from the criterion with practice, then predictive validity decreases with practice. In the latter case there is no point in extending practice on the predictor task; in the former case there may be, especially if the increase in predictive validity is sizable.

Flight grades may also converge on or diverge from the predictor task, taken, let us say, as terminal levels of performance. This is the second kind of convergence-divergence in the selection context. If flight grades diverge from the predictor task as a student progresses to more and more advanced stages of training, the task may easily be valid in training but much less so or not at all in operations. If, on the other hand, flight grades converge on the predictor task, there is reason to believe that the test will predict operational performance at least as well as it does performance in training.

Predictor-criterion relations are seriously oversimplified when presented in a static point-to-point way. In many selection programs both the predictor and the criterion change systematically with practice or training; where they do, it is crucial to know not just the overall magnitude of predictor-criterion relations but how these relations change with stage of training or practice on the predictor task.

FACTOR IDENTIFICATION

The third application concerns factor-referenced tests. It has been known for more than 20 years that the factorial content of a skilled task changes with practice (Fleishman, 1960; Fleishman and Hempel, 1953). Twenty years ago, however, it was customary to draw a sharp distinction between skills and abilities. Skills were tasks of practical relevance and they were practiced. Abilities were measured by "reference tests" which were administered for short periods of time only and usually had little or no practical importance. Skills were narrow in scope, whereas abilities were broad and enduring. To a large extent, this distinction was always arbitrary; it was only a convention that said skilled tasks could be practiced whereas reference tests could (or should) be administered to the same subjects once or twice only. As long as it lasted, however, the distinction served to contain and limit Fleishman's findings about differential changes with practice. One thought of skills as changing with practice; nothing was said about abilities.

In recent years the idea of abilities as broad, measureable, enduring traits has been called into question along several lines (Alvares and Hulin, 1972, 1973; Mischel, 1968; Humphreys, 1968). One such line is to treat tests of ability like other tasks, that is, to allow practice. Some factor-referenced tests can be practiced as they stand, others require multiple parallel forms. In any case, when practice is allowed, tests of ability behave just like other tasks. People improve with practice, usually in a negatively accelerated way; correlations between trials of practice follow the usual, so-called superdiagonal pattern (Jones, 1969, 1972), with intertrial correlations decreasing with increasing serial or temporal separation; finally, as practice proceeds, tests converge on or diverge from some external measures, including other tests.

Code substitution or digit-symbol, as it is also called, is a good example. The test is generated by pairing the nine digits (1,2,3,..., 9) with nine arbitrarily chosen letters. The letters are then presented in a random series numbering, perhaps, 100 to 200 letters in all and the subject required to write down the paired digit for as many letters as possible in the time allowed. The usual measure is number correct or time to finish, if the series is short relative to time allowed.

In one form or another this test has been included in intelligence tests since the First World War. It was part of the Army Alpha and Beta tests (Pintner, 1923) and the original version of the Wechsler-Bellevue Intelligence Scale (Wechsler, 1958). The Differential Aptitude Tests, General Aptitude Test Battery, and most tests of clerical aptitude include some form of the code substitution test (Buros, 1972).

Recently, Pepper, Kennedy, and Bittner (1980) administered alternate forms of the code substitution test to 19 Navy enlisted men for 15 consecutive working days. Each alternate form consisted of 135 letters with its own randomly chosen letter-digit pairing. The subjects were instructed to write down the paired digit beneath each letter and given two minutes to complete as many pairs as they could. The measure of performance was total number correct.

Mean performance on this test increased regularly from 68.3 correct on the first day to 80.2 correct on the 15th day; the variance among subjects stabilized after day 7. Intertrial (interday) correlations showed unmistakable evidence of differential change for the first five days but little or none thereafter. The average correlation among the last ten days, however, was not as high as one might wish, .72 (Jones, 1979).

This study was carried out at the Navy Biodynamics Laboratory and many of the same subjects were also given other tasks, for example, Air Combat Maneuvering. The correlations of code substitution with several of these other tasks changed systematically with practice, in some cases increasing and in others decreasing.

In short, code substitution behaves just like other tasks when it is practiced, its differential content changes; and what is true of code substitution is probably true of most tests, including tests used to identify latent factors. This fact poses serious problems for factor analysis.

Suppose, for example, that code substitution loads heavily on factor A when it is little practiced but only very modestly when it is well practiced. Given the first result, the usual conclusion would be that factor A had something to do with clerical aptitude, speed, or accuracy. But if that is true, then why doesn't code substitution load heavily on factor A when it is well practiced? If factor loadings change with practice--and this much is foregone given that tests converge on or diverge from one another with practice, then how are ability factors to be named? The same test that loads heavily on a factor at one stage of practice may not do so at another; yet the content of the test, its behavioral requirements, remains the same.

One way out of this dilemma is to equate test content with terminal levels of performance. On this view, early stages of skill acquisition reflect previous experience, differences in exposure, variations in learning style, etc.; it is only late in practice, when subjects approach asymptotic performance, that one can say, "these differences reflect test content." This view also entails difficulties, however. Taken seriously, it means that most factor-analytic attempts to identify underlying abilities are improperly done since very few of them involve extended practice on any task.

REFERENCES

Adams, J.A. The prediction of performance at advanced stages of training in a complex psychomotor task. USAF Human Resources Research Center, 1953, Research Bulletin No. 53-49.

Alvares, K.M. and Hulin, C.L. Two explanations of temporal changes in ability-skill relationships: A literature review and theoretical analysis. Human Factors, 1972, 14, 295-308.

Alvares, K.M. and Hulin, C.L. An experimental evaluation of a temporal decay in the prediction of performance. Organizational Behavior and Human Performance, 1973, 9, 169-185.

Buros, O. Seventh mental measurement yearbook. Highland Park, New Jersey: Gryphon Press, 1972.

Fleishman, E.A. Abilities at different stages of practice in rotary pursuit performance. Journal of Experimental Psychology, 1960, 60, 162-171.

Fleishman, E.A., and Hempol, W.E., Jr. Changes in factor structure of a complex psychomotor test as a function of practice. USAF Human Resources Research Center, 1953, Research Bulletin No. 53-68.

Humphreys, L.G. The fleeting nature of the prediction of college academic success. Journal of Educational Psychology, 1968, 59, 375-380.

Jones, M.B. A two-process theory of individual differences in motor learning. Psychological Review, 1970, 77, 353-360. (a)

Jones, M.B. Rate and terminal processes in skill acquisition. American Journal of Psychology, 1970, 83, 222-236. (b)

Jones, M.B. Differential processes in acquisition. In E.A. Bilodeau and I. Med. Bilodeau (Eds.). Principles of skill acquisition. New York: Academic Press, 1969.

Jones, M.B. Individual differences. In R.N. Singer (Ed.), The psychomotor domain. Philadelphia: Lee and Febiger, 1972.

Jones, M.B., Kennedy, R.S., and Bittner, A.C., Jr. Video games and convergence or divergence with practice. Presented at the 7th Psychology in the Department of Defense Symposium, April 16-18, 1980, USAF Academy, Colorado Springs, Colorado.

Mischel, W. Personality and assessment. New York: Wiley, 1968.

Pepper, R.L., Kennedy, R.S. and Bittner, A.C., Jr. Development of performance evaluation tests for environmental research (PETER): Code substitution test. Presented at the 7th Psychology in the Department of Defense Symposium, April 16-18, 1980, USAF Academy, Colorado Springs, Colorado Springs, Colorado.

Pintner, R. Intelligence testing. New York: Henry Holt and Co., 1923.

Reynolds, B. Correlation between two psychomotor tasks as a function of distribution of practice on the first. Journal of Experimental Psychology, 1952, 43, 341-348.

Wechsler, D. The measurement and appraisal of adult intelligence. Baltimore: Williamson and Wilson, 1958.

Woodrow, H. Factors in improvement with practice. Journal of Psychology, 1939, 7, 55-70.

PROCEEDINGS OF THE 23RD ANNUAL MEETING OF THE HUMAN FACTORS SOCIETY
BOSTON, OCTOBER, 1979

STATISTICAL TESTS FOR DIFFERENTIAL STABILITY

Alvah C. Bittner, Jr.

Naval Aerospace Medical Research Laboratory Detachment
New Orleans, Louisiana

ABSTRACT

This paper evaluates three methods for assessing "differential stability" These methods are Graphical Analysis, Early vs. Late Correlational ANOVA, and the Lawley Test of Correlational Equality. It is recommended that Graphical Analysis be the method of first choice with the Early vs. Late method utilized only where there is a need for formal confirmation.

INTRODUCTION

Background

Development of Performance Evaluation Tests for Environmental Research (PETER) is currently taking place at a number of government, university and industrial facilities. PETER is a human performance task battery which is being specifically designed for repeated administration in exotic environments. Focus on repeated administrations was motivated by recognition that the most frequently and almost exclusively used paradigm in environmental research uses repeated measurements of subjects. With and without control groups, this paradigm typically employs measurements of subjects in "before", "during" and "after" exposure conditions. Suitability of tasks for repeated administrations is a unique focus of PETER not considered in previous battery developments (Kennedy & Bittner, 1978; Kennedy, Bittner & Harbeson, 1979).

The repeated measures analysis of variance (ANOVA), almost universally applied to environmental paradigm data, puts stringent requirements on tasks for use in PETER. One of the requirements of such an ANOVA is "compound symmetry" of the variance-covariance matrix (Winer, 1962). This requirement can be shown (Anderson, 1958) equivalent to requiring: (a) homogenity of variances across conditions and (b) differential stability, i.e., the correlations between trials must be constant ($\rho_{ij} = \rho_{ij}$, $i \neq j$). Usually the first of these requirements, homogenity of variance, can be met by either differential weighting of observations or transformations (Scheffe', 1959). Hence differential stability is the critical assumption for conventional analysis of environmental paradigm data.*

Differential stability, in light of its experimental importance, has been surprisingly little studied. However, a few researchers (e.g., Fleishman, 1967) have shown instability for some tasks by demonstration of systematic variations in correlations between a reference battery and trial-to-trial performance on a task. In addition, the decline in between-trial correlations (sometimes to zero) with increasing trial separation has been noted by Jones (cf, 1962 and 1972) and followers (Kennedy & Bittner, 1978b) to suggest differential instability for almost all tasks without extensive practice. Kennedy and Bittner (1978b), in their study of potential tasks for PETER, have noted differential instability even where mean and standard deviations have "plateaued" and most experimenters would assume sufficient "stability" for conduct of research. More recent PETER investigations have also found many tasks differentially "unstable" after practice ordinarly thought sufficient for their experimental utilization (Kennedy, Bittner & Harbeson, 1979). Clearly, there is need of methods for assessing if and when tasks obtain differential stability.

Purpose

The purpose of this report will be to evaluate three methods of assessing differential stability.

TESTS OF STABILITY

Three tests of differential stability will be described below: (1) Graphical; (2) Early versus Late Correlational ANOVA; and (3) Lawley (1963) Test of Correlation Equality. Each of these tests will be illustrated by using the between trial correlations obtained from thirteen subjects who practiced a video game, ATARI Air Combat Maneuvering, for 10 trials a day over 15 days. Table 1 gives the ATARI correlation matrix which has been des-

*Multivariate profile analysis of basic environmental paradigm data can be conducted, despite the lack of differential stability (or homogenity of variances), if the number of subjects exceeds the number of trials. Lack of differential stability, however, implies that the character of what-is-being-measured, is not constant over trials. Hence, while statistically valid, multivariate analysis may yield results meaningless from a scientific stand point.

cribed elsewhere (Jones, Kennedy & Bittner, 1979).

Graphical Analysis

Studies of differential stability by Graphical Analysis have been reported by Kennedy and Bittner (1978 a&b) and Kennedy et.al. (1979). While not yielding a strictly statistical test, Graphical Analysis permits visual understanding of task progression toward and attainment of stability, if present. Consider Figure 1 which portrays the correlations between selected base days (1, 2, 4, 6, 10 & 12) and those which follow. It was constructed by selecting a row of Table 1 corresponding to a base day of interest (e.g., Day 2) and plotting the correlations to the right of the diagonal in terms of "Days After Base Day (DABD)" i.e., (r_{23} = .92 at 1 DABD, r_{24} = .87 at 2 DABD). Differential stability can be determined from the traces such as portrayed in Figure 1, by noting where the slopes of later Base Day traces approximate zero and overlay one another. A zero flat slope, it is noteworthy, indicates that correlations are stable in value and the overlay of traces indicates that correlations are equal across Base Days. Examining Figure 1, traces for Base Days 1, 2 and 4 are seen to lie below a cluster of later Days and to have apparently negative slopes. Traces for Base Days 6 and later, however, appear to effectively overlay one another and have zero slopes. From Graphical Analysis, therefore, it appears that differential stability has been obtained on the ATARI task by the sixth day of practice.

Early vs Late Days Correlation ANOVA

Jones (1979) has defined and applied this method of stability analysis. Following Jones it can be argued that if stabilization occurs, the practice days can be divided into an "earlier" and a "later" segment such that: (a) the correlations between all of the later days and one of the earlier days is constant and (b) the correlation between any two later days is the same. This Early vs. Late days division, Jones (1979) observes, can be seen in examination of a table of cross-correlations and subjected to ANOVA.

Delineation of Jones (1979) method of analysis can be made with the ATARI data given in Table 1. Consider Table 2 which presents the correlations between the first six (tentatively early) and the last nine (tentatively "late" days. The rows subject to sampling variation, appear to meet the first (a) of Jones conditions for stability with relative consistancy going across any row. The average correlations for the columns present support for meeting the second (b) of Jones conditions as there appears to be no change at all from Day 7 to Day 15. In other words, Day 7 correlates no more strongly with the first six days than Day 15 does. It appears, therefore, that the ACM task is completely stabilized after Day 6. Table 3 summarizes the results of a two way analysis of variance (ANOVA) carried out on the correlations in Table 2. Only the linear columns component is of interest because it reflects the flatness of early correlations with later days. Being nonsignificant (F=1.0, p>.6), the tentative interpretation of stability of correlation after Day 6 is confirmed.**

Lawley Test of Correlation Equality

Lawley (1963) has proposed a test for the equality of all correlations in a matrix, i.e., $H_0: \rho_{ij} = \rho$ $(i \neq j)$. His test, is an approximation of a likelihood-ratio test and rests on the assumption that the underlying distribution of observations is multivariate normal. Lawley's test statistic (Morrison, 1967) can be written

$$\chi^2 = \frac{n}{\hat{\lambda}^2}\left[\sum_{i<j}\sum (r_{ij} - \bar{r})^2 - \hat{\mu}\sum_{k=1}^{p}(\bar{r}_k - \bar{r})^2\right]$$

where for p variates and N subjects

$$n = N - 1$$

$$\hat{\lambda} = 1 - \bar{r}$$

$$\hat{\mu} = \frac{(p-1)^2(1-\hat{\lambda}^2)}{p - (p-2)\hat{\lambda}^2}$$

$$\bar{r}_k = \frac{1}{p-1}\sum_{\substack{i=1 \\ i \neq k}}^{p} r_{ik}$$

$$\bar{r} = \frac{2}{p(p-1)}\sum_{i<j}\sum r_{ij}$$

Under the assumption H_0, Lawley (1963) has shown that asymptotically his test statistic is chi-squared distributed with df=½(P+1)(P-2) degrees of freedom. Applying the Lawley statistic to the 36 correlations among days 7 through 15 of Table 1, it can be found that the chi-squared is 39.82 which for 35 degrees of freedom is nonsignificant (p>.75). Hence the conclusion of differential stability subsequent to the sixth day is again confirmed.

**Non linear column effects, it is noteworthy, are not of interest as they largely reflect non systematic sampling variations. In this case, the spuriously low average correlations on Day 10, a Friday, is largely responsible for the nonlinear effect.

DISCUSSION

Each of the three stability tests was found to indicate differentially stability for the ATARI ACM task by Day 7. This concensus, however, masked important differences between the three methods. These differences will be described below and recommendations will be made for statistical method selection.

Test Differences

Recently, Jones (1979) has pointed out that the Early vs Late Days and Lawley methods examine stability differently. The Lawley test has its focus on the equality of all correlations within a series of consecutive trials. Therefore, it can be expected to be sensitive to local deviation in correlations, reflecting more accidental disruptions in performance than changes in differential stability (e.g., an unscheduled break during testing for some subjects). The Early vs Late test, in contrast, has its focus on systematic (linear) changes in average correlations with an external criteria (early trials). Local instabilities, effecting the Lawley, would be expected to have little impact on the Early vs Late method. Jones (1979) has defined the stability measured by the Lawley as local and that by the Early vs Late as general. In light of Jones distinction, Graphical Analysis can be seen to focus on "general" stability paralleling the Early vs Late method.

Each of the three stability methods can be distinguished by "cautions" for the potential user. Graphical Analysis in particular, is not, strictly speaking, an objective statistical technique. It does not yield an alpha level or other numerical assessment. Interpretation of graphical traces requires a "knowledgeable eye" and disagreements between analysis, although infrequent, are possible.

The Early vs Late Days Correlational ANOVA is more objective than the Graphical technique, but the Early vs Late Days test statistic may yield significance levels which are substantially in error. An arguement to show this possibility can be made from the observation that elements in estimated covariance matrices have correlated errors (cf Anderson, 1958). Consequently, correlations estimated from covariance matrix elements will also have correlated errors, errors which might be expected to impact significance levels at a substantial level if experience with lag correlated errors is any indication (Scheffe', 1959, Chap. X). It can be noted, however, that analysis has suggested that the impact of correlated errors for matrices arising from reliability studies will be to inflate the apparent significance level. Hence a nonsignificant (linear column) result for the Early vs Late Days Analysis would support the view that a task is stable.

The Lawley Test, as with both the other methods described above, must be used cautiously by researchers. It is based on an assumption of multivariate normality which if violated could yield grossly inappropriate estimates of alpha level. An arguement for this sensitivity to nonnormality can be constructed following that for the sensitivity of tests for homogenity of variance (e.g., Bartlett) given in Scheffe' (1959). Thus the user of the Lawley Test must attend to the multivariate distribution underlying observations.

Recommendations

One goal of stability research is to determine if differential stability is sufficient for utilization of a task in an exotic environment. For many tasks, Graphical Analysis alone is sufficiently precise to meet this goal. In cases of massive declines in reliability (e.g., McCauley, et.al., 1979), the task can be rejected without resort to more elegant techniques. In other cases (e.g., Seales, et.al., 1979), the graphical evidence for stability is so marked that evidence from the Early vs Late and Lawley Tests could be discounted as meaningless from a practical standpoint. Even in cases where stability or instability is difficult to assess (e.g., Kennedy & Bittner, 1978a), Graphical Analysis is sufficiently precise to indicate sufficient (practical) stability for task use in a limited number of test periods. Because of the wide utility and simplicity of Graphical Analysis, it is suggested as the first step in stability analysis. Reliance on a non graphical method can be confined to situations where graphical analysis is inconclusive. In cases where confirmation is required, Early vs Late Days appears the current method of choice. It measures "general stability" which is more practically meaningful than "local stability" assessed by the Lawley. Hence, Graphical Analysis is the recommended method of first choice with Early vs Late Days Analysis recommended only where a special need for confirmation manifests itself.

REFERENCES

Anderson, T. W. Introduction to Multivariate Statistical Analysis. New York: Wiley, 1958.

Fleishman, E. A. "Individual differences and motor learning" In R. Gagne' (Ed.) Learning and Individual Differences. Columbus: Merrill, 1967 pp. 165-191.

Jones, M. B. Practice as a process of simplification. Psycho. Reviews, 1969, 69, 274-294.

Jones, M. B. "Differential processes in acquisition" In E. A. Bilodeau & I. McD. Bilodeau (Eds.) Principles of Skill Acquisition. New York: Academic Press, 1969.

Jones, M. B. "Rate and terminal process in skill acquisition". American Journal of Psychology, 1970, 83, 222-236.

Jones, M. B. Individual differences. In R. N. Singer (Ed.), The Psychomotor Domain. Philadelphia: Lea and Febiger, 1972.

Jones, M. B. "Stabilization and task definition in a performance test battery" Pennsylvania State University College of Medicine, Unpublished Final Report on Contract N0023-79-M-5089, 12 May 1979 (b) (Copies available from NAMRLD, New Orleans, LA).

Jones, M. B., Kennedy, R. S., & Bittner Jr., A. C., "A video game for performance testing: Paper presented at Rocky Mountain Psychological Association Annual Meeting, Los Vegas, NV, May, 1979. (Paper available from NAMRLD, New Orleans, LA).

Kennedy, R. S., & Bittner Jr., A. C. "The development of a Performance Evaluation Test for Environmental Research (PETER)". In Productivity Enhancement in Navy Systems, San Diego: Naval Personnel Research and Development Center, October, 1977. AD#A056047.

Kennedy, R. S., & Bittner Jr., A. C. "The stability of complex human performance for extended periods: applications for studies of environmental stress. Presented at Aerospace Medical Association, New Orleans, LA, May 1976 . (Printed in Preprints). (a)

Kennedy, R. S., & Bittner Jr., A. C. "Progress in the analysis of a Performance Evaluation Test for Environmental Research (PETER)." Proceedings 22nd Annual Meeting of the Human Factors Society, Detroit, MI: HFS, 1978. (b) AD#A060676.

Kennedy, R. S., Bittner Jr., A. C., & Harbeson, M. "Research developments in a Performance Evaluation Test for Environmental Research (PETER)." Presented at the Undersea Medical Society Annual Scientific Meeting, Key Biscayne, FL, May 1979. (Paper available from NAMRLD, New Orleans, LA).

Lawley, D. N. "On testing a set of correlation coefficients for equality", Annals of Math. Stat., 1963, 34, 149-151.

McCauley, M. E., Kennedy, R. S., & Bittner Jr., A. C. "Development of Performance Evaluation Tests for Environmental Research (PETER): Time Estimation Tests" in Proceedings of the 23rd Annual Meeting of the Human Factors Society, Boston: HFS, October, 1979.

Morrison, D. F. Multivariate Statistical Methods. New York: McGraw-Hill, 1967.

Scheffe', H. The Analysis of Variance. New York: Wiley, 1959.

Seales, D. M., Kennedy, R. S., & Bittner Jr., A. C. "Development of Performance Evaluation Tests for Environmental Research (PETER): Arithmetic Computation". Proceedings of the 23rd Annual Meeting of the Human Factors Society, Boston: HFS, October, 1979.

Winer, B. J. Statistical Principles in Experimental Design. New York: McGraw-Hill, 1962.

FIGURE

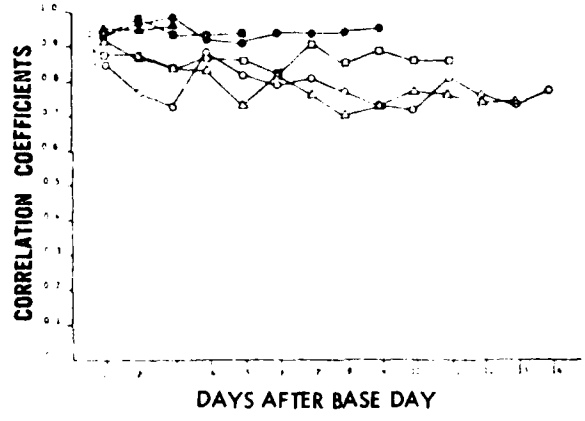


Figure 1. Comparison of reliabilities for selected Base Days (1,2,4,6,10 & 12).

TABLES

Table 1
Correlations Among Days for Atari ACM Task

Day	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	-	.85	.77	.73	.88	.82	.79	.81	.77	.73	.72	.81	.76	.73	.77
2		-	.92	.87	.84	.83	.73	.82	.76	.70	.73	.77	.76	.74	.74
3			-	.90	.88	.84	.73	.80	.81	.70	.81	.73	.79	.74	.78
4				-	.88	.88	.84	.87	.86	.82	.91	.85	.89	.86	.86
5					-	.95	.91	.95	.94	.90	.93	.91	.93	.89	.92
6						-	.93	.97	.98	.92	.91	.94	.94	.94	.95
7							-	.97	.92	.93	.94	.96	.94	.93	.96
8								-	.95	.95	.93	.97	.94	.94	.96
9									-	.92	.94	.93	.94	.94	.94
10										-	.93	.98	.94	.93	.94
11											-	.93	.96	.94	.95
12												-	.95	.95	.96
13													-	.98	.98
14														-	.97
15															-

Table 2
Correlations Between First 6 and Last 9 Days

	Last Nine Days									
First Six Days	7	8	9	10	11	12	13	14	15	$\bar{r}$
1	.79	.81	.77	.73	.72	.81	.76	.73	.77	.766
2	.73	.82	.76	.70	.73	.77	.76	.74	.74	.750
3	.73	.80	.81	.70	.81	.73	.79	.74	.78	.766
4	.84	.87	.86	.82	.91	.85	.89	.86	.86	.862
5	.91	.95	.94	.90	.93	.91	.93	.89	.92	.920
6	.93	.97	.98	.92	.91	.94	.94	.94	.95	.942
$\bar{r}$	.822	.870	.853	.795	.835	.835	.845	.817	.837	.834

Table 3
ANOVA For Data in Table 2

Source	SS	df	MS	F	P
Rows	0.3253	5	0.0651	108.5	<.001
Columns	0.0225	8	0.0028	4.8	<.001
linear	0.0006	1	0.0006	1.0	n.s.
residual	0.0219	7	0.0031	5.2	<.001
Interaction (error)	0.0233	40	0.0006		
Total	0.3711	53			

PROCEEDINGS OF THE 51ST ANNUAL MEETING OF THE AEROSPACE MEDICAL ASSOCIATION
ANAHEIM, CA 12-15 MAY 1980

PHYSIOLOGICAL AND PERFORMANCE MEASUREMENTS: A TIME-SERIES MODEL

Robert C. Carter
Naval Aerospace Medical Research Laboratory Detachment, New Orleans, LA 70189

Some of the most interesting phenomena of psychology, physiology, and medicine develop over time. Investigators of these dynamic phenomena suggest that the best way to study them is to measure an individual repeatedly, and to gain generalizability by studying several individuals. For example, Hecht, Haig, and Chase(5) studied individual dark adaptation curves because composite curves obscure the premier feature of adaptation: the rod-cone break. Similarly, Estes(4) showed that learning curves based on group data misrepresent learning by individuals. More recently, Klien and Armitage(7) demonstrated 90-minute oscillations of mental abilities which would be obscured by averaged performance curves and classified as error variance by traditional statistical analyses.

Data such as these are in the form of a series of observations separated by equal intervals of time (a time series), in which each observation depends on those which precede it. Traditional methods of data analysis are inadequate for these kinds of data because "ordinary parametric or nonparametric statistical procedures which rely on independence or special symmetry in the distribution function are not available nor are the blessings endowed by randomization"(2).

In response to this dilemma, Box, Jenkins and their colleagues have recently developed a system for time series analysis(1). Their model is similar to the psychological model: S→O→R, in which a series of Stimuli (S) cause an Organism (O) to produce a series of Responses (R). In the Box-Jenkins model, the Stimuli at times t are called "Input$_t$", the Organism is called a "Transfer Function", and the responses are called "Output$_t$". The organism's response-time and memory are represented by delays in the transfer function (d_i is a delay of i epochs, see Figure 1).

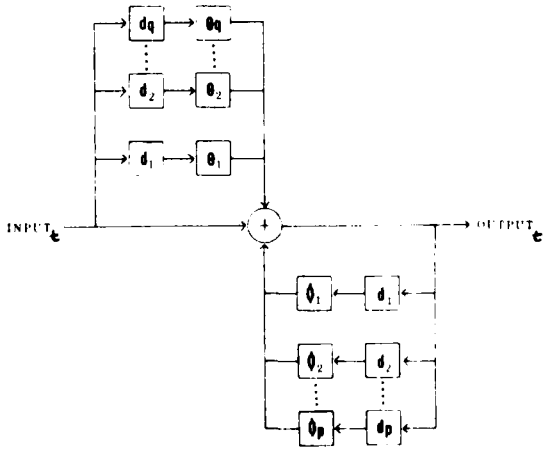


Figure 1. Block Diagram of the Box-Jenkins Model

The influence of past inputs and outputs on present output is proportional to the model's parameters (θ and ϕ's).

The Box-Jenkins model is a dynamic, stochastic, discrete representation which offers the following to students of performance and physiological variables (PPV): 1) Insight into how PPV change over time, including estimates of their differential equations(1); 2) Identification of rhythms and periodicities of PPV and phase relations among PPV(1); 3) Reduction of error variance by explaining some of that variance as covariance among observations(1); 4) Explanation of how PPV (e.g. performance test scores) change in response to other PPV (e.g. vibration exposure history)(1); 5) Dynamic forecasts of PPV, including point estimates and confidence intervals which change appropriately for each future time(8); and 6) Assessment of whether some intervention(2) (e.g. clinical or environmental) affects the level of a PPV. Both univariate(1) and multivariate(9) models are available for each of these objectives. The general applicability of the Box-Jenkins model to behavioral phenomena is illustrated by the fact that a simple Box-Jenkins-type model(3) explains the simplex matrix of intertrial correlations, which characterizes all known repeated-measures data.(6)

Some of the uses of the Box-Jenkins time series model can be exemplified with data on tests of arithmetic ability collected at 6 A.M., 2 P.M., and 10 P.M. on each of 21 successive days. Models of the obtained performance were built using procedures described by Box and Jenkins(1). The primary basis of such models is the correlations between observations separated by a fixed number of measurements: autocorrelations. Figure 2 shows the autocorrelations of addition tests.

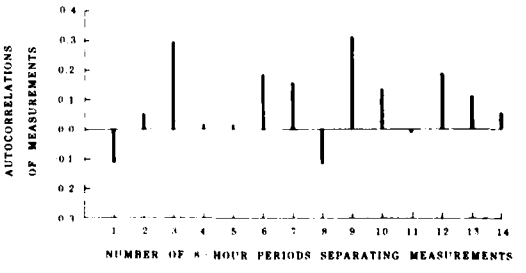


Figure 2. Autocorrelations among repeated measurements of addition.

It indicates that there is a relationship between scores obtained at 24-hour intervals. The nature of this 24-hour cycle is that perfomance at 2 P.M. was usually poorer than performance at 6 A.M. or 10 P.M. The same 24-hour cycle was discovered

in subtraction, multiplication, and division test performance. For instance, models of addition and subtraction performance are, respectively, $z_t = .326z_{t-3}+29.569$ and $z_t = .477z_{t-3}+42.33$, where z_t is the number of arithmetic problems worked correctly during the t^{th} four-minute trial of the experiment. All coefficients in these models are statistically significant ($p < .05$), and the subtraction model, for example, reduces the error variance of that series by 21%. A Chi-squared test for residual autocorrelation in the modeled series indicates that the addition and subtraction models are complete, $\chi^2(24)=9.35$, $p > .5$; and $\chi^2(24)=25.04$, $p > .3$ respectively. Such models may be used for description of a process, for intervention analysis, or for forecasting.

Dynamic forecasts of addition performance (scaled to have a mean of 50) are shown in Figure 3. A separate forecast is generated for each time in the future. Forecasts of the distant

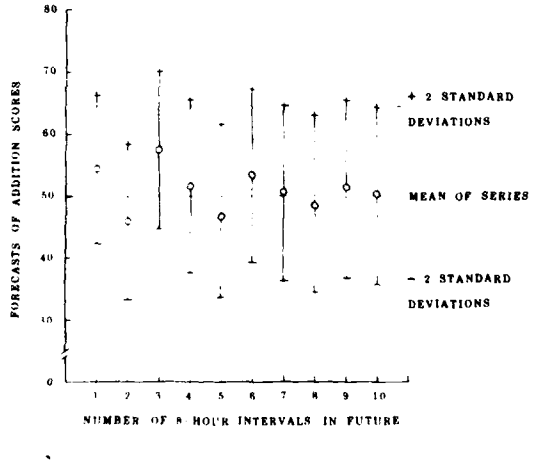


Figure 3. Dynamic Forecasts of Addition Scores, with 95% Confidence Intervals.

future approach the series mean, and their variances approach the variance of the series. Forecasts of the near future differ appreciably from the series mean, and have reduced variance due to the covariance between the near future and the (now certain) past. Note that traditional 95% confidence intervals (± 2 S.D.) will often be too liberal or too restrictive, compared with confidence intervals based on a Box-Jenkins model. Dynamic forecasting has obvious applications to manpower planning and selection. An application to aerospace medicine would be the comparison of observed PPV scores with predicted scores to indicate the incipient disability of critical personnel (e.g. aircraft pilots).

To summarize, Box-Jenkins time series models deserve our consideration as an aid to understanding, prediction, and control of psychological and physiological processes which unfold over time. These dynamic models represent a departure from traditional static models and their adoption would require a shift to experimental designs that include measurement of a few individuals on numerous occasions.

REFERENCES

1. Box, G. E. P., and Jenkins, G. M. 1970. Time series analysis forecasting and control. Holden-Day, San Francisco.
2. Box, G. E. P., and Tiao, G. C. 1975. Intervention analysis with applications to economic and environmental problem. Journal of the American Statistical Association, 70: 70-79.
3. Corballis, M. D. 1965. Practice and the simplex. Psychological Review. 72, 399-406.
4. Estes, W. K. 1956. The problem of inference from curves based on group data. Psychological Bulletin. 55: 134.
5. Hecht, S., Haig, C., and Chase, A. M. 1937. The influence of light adaptation on subsequent dark adaptation of the eye. Journal of General Physiology. 20: 831-850.
6. Jones, M. B. 1966. Individual Differences. In E. A. Bilodeau (Ed.) Acquisition of skill. Academic Press, New York: pp. 109-146.
7. Klien, R., and Armitage, R. 1979. Rhythms in human performance: 1½-hour oscillations in cognitive style. Science. 204, 1326-1328.
8. Nelson, C. R. 1973. Applied time series analysis for managerial forecasting. Holden-Day, San Francisco.
9. Tiao, G. C., and Box, G. E. P. 1979. An Introduction to Applied Multiple time series analysis. Technical Report No. 582, Department of Statistics, University of Wisconsin.

IEEE INTERNATIONAL CONFERENCE ON CYBERNETICS AND SOCIETY, BOSTON, 8 - 10 OCTOBER 1980

A SIGNAL DETECTION THEORY FUNCTION AND PARADIGM FOR RELATING SENSITIVITY (d') TO STANDARD AND COMPARISON MAGNITUDES

Alvah C. Bittner, Jr.
Naval Biodynamics Laboratory, New Orleans, LA 70189

Douglas C. Chatfield
Texas Tech University, Lubbock, TX 79409

Summary

A general signal-detection theory (SDT) psychometric function is derived which relates both comparison (ϕ_i) and standard (ϕ_j) stimulus magnitudes to sensitivity (d'_{ij}). Applicable to a breadth of stimulus dimensions, this function is

$$d'_{ij} = P_1(\ln(\phi_i + P_2) - \ln(\phi_j + P_2))$$

where P_1 and P_2 are constants. To illustrate a paradigm for identifying the P_i ($i = 1,2$), three subjects performed a lifted-weight task. Subjects made 64 judgements at each of six standards (0.1 to 1.3kg), with eight comparison weights per standard (91% to 109% of standard). The results of analyses of individual subject's data by nonlinear least squares revealed that the general model provided significantly better fit over other models ($p < 10^{-6}$) and accounted for 94% of each subject's total variation. The centroid of this model was determined to be

$$d'_{ij} = 20.32(\ln(\phi_i + 0.0785) - \ln(\phi_j + 0.0785))$$

where model parameters were the average of respective subject parameters. Comparisons of this centroid model and historical results are made. It is concluded that: the utility of functions relating sensitivity to both standard and comparison magnitudes is greater than the traditional partial expressions; and the multiple-standards-comparisons paradigm provides for a powerful comparison of psychometric functions.

Introduction

Background

Signal detection theory (SDT) has introduced decision analysis into Psychology as a model for human psychophysical behavior[7,14]. Largely based on the mathematical work of Wald[32], STD assumes that an "ideal-observer" can calculate the probabilities of an observed stimulus having been produced by a signal (plus noise) or by noise alone. These probabilities, SDT further assumes, are combined by the ideal-observer with a priori signal probabilities and decision costs (or payoffs) into a likelihood ratio "classification function". This classification function is used to optimally decide whether or not an observed stimulus contained a signal[15,24,25]. Usually making the (local) assumptions that signal and noise distributions are Gaussian with dissimiliar means and common variance, SDT identifies a sensitivity metric (d') which is mathematically invariant for differing cost and a priori probability conditions[14,25]. The utility of SDT theory lies in the approximation of the ideal-observer model to human behavior. This approximation is relatively close with human sensitivity (d') having been found to be relatively constant in studies where prior odds, payoffs, and procedures were varied. These studies have been conducted over several sensory modalities (e.g., visual and auditory) and perceptual tasks (e.g., detection and discrimination)[7,25]. Since the introduction of SDT, psychophysical researchers have largely focused on testing either the degree that the human observer acts as an ideal-observer or the effects of experimental conditions on sensitivity (d') and bias.

Researchers using SDT methodology have not concerned themselves with many of the problems of classical psychophysics[15]. With minor exceptions (e.g., Wuest[37]), they have not studied the nature of "psychometric functions" which relate judgement probabilities for stimuli when they are compared to a "standard-stimulus". In addition, SDT based researchers have not been concerned with the related problems of changes in relative observer sensitivity with changes in standards (i.e., changes in the "Weber Fraction" as a function of standard). This failure to address classical problems is, in part, the result of the conceptual paradigm usually applied by SDT researchers. In this paradigm, sensitivity (d') is assumed linear to the common measure of stimulus intensity[14,15]. For reasons suggested by Thurstone[30,31], this assumption is approximately met because any measure of intensity is (locally) linear to a scale where the assumption would be valid.* The usual SDT procedure is also not conducive to study of classical function studies because of the numbers of observations typically taken to estimate a single sensitivity. In the body of this report, a model and procedure will be described which address the classical psychophysical

* This is apparent from the Taylor's Series where $f(x + \Delta x) \approx f(x) + f'(x)\Delta x$

problems described above. Specifically, the report will be directed at the problem of determining a SDT psychometric function relating comparison (ϕ_i) and standard (ϕ_j) stimulus magnitude to sensitivity (d'_{ij}) in a detection or discrimination experiment.

Purpose

The purposes of this report are to: (1) derive a general psychometric function which relates comparison (ϕ_i) and standard (ϕ_j) magnitude to SDT sensitivity (d'_{ij}); (2) illustrate a paradigm for determining the specific form of the d'_{ij} function with data from a lifted-weight task; and (3) to demonstrate the utility of the d'_{ij} function by comparison of a centroid lifted-weight model with classical results.

A General Psychometric Function

In this section, a function relating SDT sensitivity (d'_{ij}) to comparison (ϕ_i) and standard (ϕ_j) magnitudes will be derived. This function will be shown to be

$$d'_{ij} = P_1(\ln(\phi_i + P_2) - \ln(\phi_j + P_2)) \qquad (1)$$

where the P_i ($i = 1,2$) are constants specific to a stimulus dimension. The derivation of (1) will be based on the Brentano-Ekman Law which will be described before preceeding with the derivation.

Brentano-Ekman Law

The Brentano-Ekman law is a combination of a conjecture of Brentano, 1874, and contemporary direct-scaled power laws[23,28]. Brentano's conjecture was that an increment of sensory variability in subjective units ($\Delta\psi$) is directly proportional to the stimulus in the same units (ψ), i.e.,

$$\Delta\psi/\psi = K \qquad (2)$$

where K is a constant[5]. Experimentally, $\Delta\psi$ in (2) is the amount of change in ψ which alters detection or discrimination probability Z-scores through a fixed range (e.g., $\Delta\psi$ is frequently determined for a unity change in Z, $Z = 1$). Hence, (2) can be rewritten in the form

$$\Delta\psi/\psi = k\Delta Z \qquad (3)$$

where k is a constant ($k = K/\Delta Z$). Ekman and his collaborators[3,4,9-15] are credited with establishing the generality of Brentano's conjecture when the subjective units (ψ) are linked to physical magnitudes by a direct-scaled power law[14,23,28]. The general form of this law, which will be used in derivation of (1), can be written

$$\psi = C(\phi + P_2)^B \qquad (4)$$

where C, P_2 and B are constants which vary for perceptual dimensions[12,23]. Recent studies by Teghtsoonian have indicated that, for a simplified version of (4), the Brentano-Ekman law approximately holds across more than two dozen perceptual dimensions[26,27]. Hence, the function (1) which will be derived, can be expected to have substantial generality.

Derivation

The function (1) can be derived by "integration" of (3), substitution of (4) into (3), and insertation of the result into the definition of (d') sensitivity. In particular, letting the increments $\Delta\psi$ and ΔZ become differentials in (3) and integrating,

$$kZ = \ln(\psi) + C^* \qquad (5)$$

where C* is a constant of integration. Substituting (4) into (5), it follows that

$$kZ = \ln(C(\phi + P_2)^B) + C^* \qquad (6)$$

or

$$Z = P_1(\ln(\phi + P_2) + C_*) \qquad (7)$$

where $P_1 = B/k$ and $C_* = C^* + \ln C$. To derive (1), it is necessary only to determine Z_i and Z_j for stimulus magnitudes ϕ_i and ϕ_j from (7) and substitute into the definition of sensitivity ($d' = Z_i - Z_j$). It is pertinent to note that in (1), the notation "d'_{ij}" is used to indicate functional dependency on ϕ_i and ϕ_j. Another more general derivation of (1) has been given elsewhere by Bittner[1].

A Multiple Standards-Comparison Paradigm

In this section, a Multiple Standards-Comparison Paradigm (MSCP) will be illustrated for identifying the constants in (1). First, the Method of the MSCP will be given for a lifted-weight task. The essential feature of this method lies in its procedure which secures data across several standards and comparison

stimuli. Second, the analysis of the data obtained by the MSCP procedure will be given for the data from the lifted-weight task.

Method

The apparatus, subjects and procedure of the illustrated weight judgement task experiment will be described below. A more comprehensive description of these has been made elsewhere[1].

Apparatus. Six series of weights were used. Each series consisted of a standard and eight comparison weights. Divided into two blocks of three, the weights of the standards were 0.1 kg, 0.4 kg, and 1.0 kg for the first block, and 0.3 kg, 0.7 kg, and 1.3 kg for the second block. Comparison weights were from 91% to 109% of the standards weight within a series. All weights were made from new half-pint paint cans (79 mm in diameter and 79 mm deep) fitted with lids and weighted with lead shot and cotton wads. To facilitate presentation, each weight of a series was appropriately labeled and placed on a wooden turntable. Weights and turntable were hidden from subjects view by a felt curtain through which they could reach. An adjustable chair was employed so that subjects could be seated with the elbow resting on a felt pad with the angle of the humerus being at about 45 degrees with respect to the body's trunk.

Subjects. The subjects (observers) were three (E-3) enlisted men on the staff of the Naval Biodynamics Laboratory as research volunteers. For six months prior to this study, the subjects had served in psychological experiments, but their only exposure to psychophysical judgement tasks was 300 trials training on the weight-task at 0.1, 0.6 and 1.0 kg standards two weeks prior to this study. To qualify as volunteers, the subjects had to be above the national average for Navy enlisted personnel in physical health, mental health, and intelligence. The subjects received extra compensation for participating in the research program. Each volunteer was recruited, evaluated, and employed in accordance with procedures specified in Secretary of the Navy Instruction 3900.3 and Bureau of Medicine and Surgery Instruction 3900.6. These instructions are based on voluntary consent and meet or exceed the most stringent provisions of prevailing national and international guidelines[29].

Procedure. Subjects were tested in two blocks of three days, with two weeks between blocks. During the first block, subjects were tested one day each with standards of 0.1, 0.4, and 1.0 kg with order and standard counterbalanced by Latin Square. During the second block, standards of 0.3, 0.7 and 1.3 kg were tested one day each in a similar counterbalanced manner. Eight comparison weights were judged 64 times against each standard by each subject.

After being brought to the laboratory for a series, a subject was seated so that his elbow rested on a felt pad, with the forearm directly forward of the shoulder. The subject was initially told that weights would be placed on the table in his grasp, and that he should lift the weight only to about one inch (2.54 cm) above the table, bending only the elbow while letting his elbow rest lightly on the pad. This lifting procedure eliminated variations in data due to lifting with wrist or shoulder[21]. Subsequent to lifting instructions, the subject was also informed that on one-half of the trials the comparison weights would be lighter and on one-half the comparison weights would be heavier. He was told that his job would be to judge if the second comparison weight was heavier. The replies "no" for not heavier and "yes" for heavier were used as judgement indicators. The judgement of the standard against the comparison weights commenced after instruction.

Results

The method of fitting models will be described below for the weight task. A comparison of models will be subsequently made and a centroid model will be given.

Model Fitting. After data collection, each subject's responses for each comparison weight, within a given subject-series, were first collected and the empirical probabilities of "heavier" responses determined. These probabilities were, in turn, converted to preliminary $(\hat{d}'_{ij})$ estimates by

$$\hat{d}'_{ij} = Z(P_{ij}) - Z(P_{jj}) \tag{8}$$

where $Z(P_{kj})$ is the Gaussian standard score transformation of the probability of "heavier" judgements when ϕ_k is the comparison stimulus magnitude and ϕ_j the standard. Each of three (d'_{ij}) functions given in Table 1 were then separately fit to the totality of each subject's data so as to minimize

$$S = \sum_j \sum_i \left[\hat{d}'_{ij} - (d'_{ij} + \sum_k \tilde{P}_k \delta_{kj}) \right]^2 \tag{9}$$

where the $\tilde{P}_k$ (k=1,6) are six parameters to be fit and δ_{kj} is a Kronecker delta*.

Table 1

Functions

$$\text{I: } d'_{ij} = P_1(\phi_i - \phi_j)$$
$$\text{IIA: } d'_{ij} = P_1(\ln(\phi_i) - \ln(\phi_j))$$
$$\text{IIB: } d'_{ij} = P_1(\ln(\phi_i + P_2) - \ln(\phi_j + P_2))$$

Models of the form (9) with the $\tilde{P}_k$ parameters, it is noteworthy, provide for utilizing all $Z(P_{ij})$ data in a series for estimating $Z(P_{jj})$ rather than just the empirical $Z(P_{jj})$. Statistical and empirical justifications for this procedure have been made by Bittner and colleagues[1,2].

All minimizations of (9) were accomplished using the nonlinear least squares computer program BMDP3R[8]. This program employed a stepwise Gauss-Newton (total differential) method which selects the parameter to be estimated at each step for greatest potential reduction in the residual sum-of-squares. Originally developed by Hartley, this technique has been shown to generally converge more rapidly than the unstepwise total differential method in difficult cases[19]. All minimizations, employing BMPD3R, were conducted using at least three initial estimates of parameters. These initial estimates were derived by various means, such as graphical estimates, parameters from simpler models, multivariate search, and other more "subjective" techniques. The use of several sets of initial estimates was to give assurance that minimum least squares were "global" vs. "local".

Comparison of Models. Figure 1 shows the percent remaining sums-of-squares for the subject "observers" over models I, IIA and IIB. Examining this figure, it appears that the remaining sums-of-squares are less than half as great for Model IIA than for Model I. Table 2 which presents statistical comparisons of I and IIA supports this view with each subject (observer) showing statistical significance ($p < E\text{-}7$)**. The Pearson P_λ statistic[23,33] combines the individual significance levels and indicates over all significance beyond $p < 1.1$ E-10. Examining Figure 1, it is also apparent that the residual sums-of-squares for Model IIB are substantially less than for IIA. This view is supported by the results reported in Table 3 where the least significant result is for Observer 1 ($p < .005$). The P_λ statistic indicates that over the subjects, the significance of this difference between IIA and IIB is beyond $p < 1.5$ E-7. Overall, the model IIB has provided significantly better fit than other models and accounted for 94% of each subject's total variation.

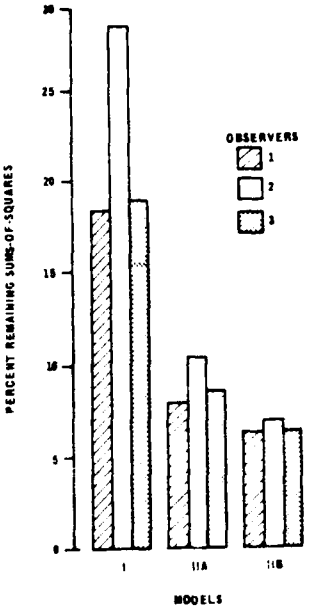


Figure 1. Percent Remaining Sum-Of-Squares for Three Models.

Table 2

Summary of Conservative Comparisons of Models I and IIA with Standard Parameters**

Observers			Combined (P_λ)
1	2	3	
$F(1, 36) = 46.918$	$F(1, 37) = 63.795$	$F(1, 37) = 44.769$	$\chi^2(6) = 107.401$
$p = 45E\text{-}9$	$p = 1.4E\text{-}9$	$p = 73E\text{-}9$	$p < 1.1E\text{-}10$

*1E-n = 10^{-n}

Table 3

Summary of Comparisons of Models IIA and IIB with Standard Parameters**

Observers			Combined (P_λ)
1	2	3	
$F(1, 35) = 8.871$	$F(1, 36) = 18.782$	$F(1, 36) = 12.545$	$\chi^2(6) = 42.438$
$p = 0.005$	$p = 1.1E\text{-}4$	$p = 0.001$	$p < 1.5E\text{-}7$

*1E-n = 10^{-n}

* $\delta_{kj} = \begin{cases} 1 & k = j \\ 0 & k \neq j \end{cases}$ ** E-n=10^{-n}

Centroid Model

Table 4 gives the values of the parameters P_1 and P_2 for each of the three observers determined by fitting IIB.

Table 4

Parameter Values for Model IIB

OBSERVER	PARAMETERS	
	P_1	P_2
1	17.17	0.0690
2	25.56	0.0822
3	18.23	0.0844

Based on the averages of the parameters, the centroid observer model is

$$d'_{ij} = 20.32\ (\ln(\phi_i + 0.0785) - \ln(\phi_j + 0.0785)) \quad (10)$$

This model is a more appropriate representation of behavior than averaged subject performance at different levels because it preserves the form of individual functions[1].

Discussion

The theoretical and practical utility of the d'_{ij} function and MSCP will be discussed in this section. Subsequent to a brief review of generality, MSCP sensitivity and d'_{ij} historical-results comparability will be delineated. Conclusions will be made based on the review and delineations.

Generality of the Derived Function

The general SDT psychometric function (1) derived earlier can be expected to characterize a breadth of stimulus dimensions because of its basis on the Brentano-Ekman law. The results of Teghtsoonian, in particular, suggest the applicability of (1) to more than two dozen sensory dimensions[26,27]. Using the MSCP, as illustrated for the weight task, the parameters of this function can also be identified. Hence, application of results of this report can be expected to yield d'_{ij} functions which will successfully characterize a wide range of stimulus dimensions.

MSCP Sensitivity

Psychometric law comparisons have frequently contrasted Fechnerian[16] phi-gamma and Thurstonian[30,31] phi-log-gamma hypotheses. These comparisons have classically been made with data obtained from a single standard and a set of comparison stimuli[18,35,36]. With a restricted range of stimuli, "... it (has) consequently been difficult to distinguish between ... hypothesis empirically..."[14] The MSCP, with a greater range of stimuli, offers greater sensitivity than the classical paradigm. This can be seen by noting the strength of the comparisons of Model I, IIA, and IIB as seen in Figure 1 and Tables 2 and 3. Viewable as analogous to the phi-log-gamma hypothesis, Model IIA was seen to have less than half the residual sums-of-squares as Model I which is similarly analogous to the phi-gamma hypothesis. For each of the three observers, this difference was highly significant ($p < E\text{-}7$) and, across observers this difference was very highly significant ($p < 1.1E\text{-}10$). In addition, Model IIB which is analogous to a generalized phi-log-gamma hypothesis[20] was found to have 20% to 30% less residual than Model IIA, and across observers this result was also very highly significant ($p < 1.5E\text{-}7$). The MSCP offers considerable sensitivity for comparison of psychometric functions.

Centroid Model and Historical Results

The centroid model (10) contains similar information to that contained in a large body of classical results: (a) body of Weber Fraction Results; and (b) Brown's Single Observer Results.

Weber-Fraction Results. Figure 2 shows Weber-Fraction (σ_ϕ/ϕ) results obtained by Fechner[16], Brown[6], Woodrow[33], Oberlin[21], and an exercise of Model (10).

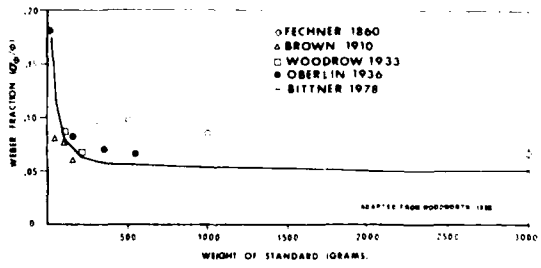


Figure 2. Comparison of Empirically Established Poikilitic Model from Current Study with Classical Results.

Examining Figure 2, it can be seen that (10) follows the body of classical results. In particular, the model is seen to be virtually on top of the results

of Oberlin from 0.025 kg to about 0.1 kg. From 0.1 kg to 0.2 kg, the model overlays Woodrow's findings. From 0.2 to 0.6 kg, the model results are seen to parallel Oberlin's and Fechner's with the paralleling of Fechner extending to 3.0 kg. Hence, in addition to results of this investigation, the centroid model (10) represents a body of Weber-Fraction results from previous investigations.

Brown's Single Observer Data. Figure 3 compares data collected by Brown[6] with centroid model (10) estimates adjusted for sensitivity.

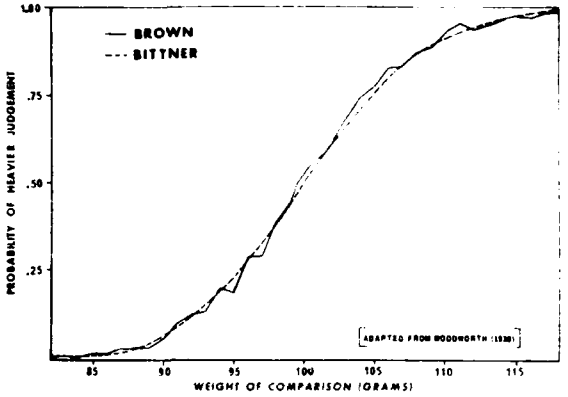


Figure 3. Comparison of Brown and Adjusted Centroid Model Results.

The Brown results were obtained in a two-category weight-lifting experiment with a standard at 0.1 kg and comparison stimuli ranging from 0.092 to 0.118 kg in 0.001 kg increments. Employing the method of constant stimuli with 700 trials at each comparison value, Brown's results are the most comprehensive study of any single psychometric function in the literature[35,30]. The model (10) was adjusted by multiplying all sensitivity estimates by 1.15 as suggested by differences between the centroid model and Brown's Weber-Fraction results seen in Figure 2. On examaination of Figure 3, the near overlay of a reasonably adjusted centroid model and Brown data is seen.

Conclusions

It can be concluded that: (a) the d'_{ij} function has wide potential for description of sensitivity across sensory dimensions; (b) the utility of the d'_{ij} function relating sensitivity to both standard and comparison is greater than traditional partial expressions; and (c) the MSCP provides for more sensitivity in comparing hypothetical psychometric functions than traditional paradigms.

References

1. Bittner, Jr., A. C. Power law poikilitic functions: Empirical test for lifted weight (Doctoral dissertation, Texas Tech University, 1978). Dissertation Abstracts International, 1979, 40/03. (University Microfilms No. 79-20,365).
2. Bittner, Jr., A. C., Kennedy, R. S., & McCauley, M. E. Time estimation: Repeated measures testing and drug effects. Proceedings of the Seventh Psychology in the DOD Symposium, Colorado Springs: USAF Academy, April 1980.
3. Bjorkman, M. Some relationships between psychophysical parameters. Red. Psychol. Lab, Univ. Stockholm, 1958, No. 65.
4. Bjorkman, M. Variability data and direct quantitative judgment for scaling subjective magnitude. Rep. Psychol. Lab., Univ. Stockholm, 1960, No. 78.
5. Brentano, F. Psychologie Vom Empirischen Standpunkt. Th.1. Leipzig: Dunker & Humblt, 1874.
6. Brown, W. The judgment of difference. Calif. Univ. Pub. Psych. 1910, 5, 115-134.
7. Coombs, C. H., Dawes, R. M. & Tversky, A. Mathematical Psychology - An Elementary Introduction. Englewood Cliffs, N.J.: Prentice Hall, 1970.
8. Dixon, W. J. & Brown, M. B. (eds.). BMDP-77 Biomedical Computer Program P-Series. Los Angeles: University of Southern California Press, 1977.
9. Eisler, H. On the problem of category scales in psychophysics. Scand. J. of Psychol., 1962, 3, 81-87 (a).
10. Eisler, H. Empirical test of a model relating magnitude and category scales. Scand. J. of Psychol., 1962, 88-89 (b).
11. Ekman, G. Discriminal sensitivity on the subjective continuum. Acta Psychologica, 1956, 12, 233-243.
12. Ekman, G. Weber's law and related functions. J. of Psychol., 1959, 47, 343-352.
13. Ekman, G. & Kunnapas, T. Subjective dispersion and the Weber fraction. Rep. Psychol. Lab., Univ. Stockholm, 1957, No. 41.
14. Engen, T. Psychophysics: Discrimination and detection. In J.W. Kling and L.A. Riggs (eds.) Experimental Psychology. San Francisco: Holt, Rinehart and Winston, 1971, 11-46 (a).
15. Engen, T. Psychophysics: Scaling methods. In J.W. Kling and L.A. Riggs (eds.) Experimental Psychology. San Francisco: Holt, Rinehart and Winston, 1971, 47-86 (b).
16. Fechner, G. T. Elements Der Psychophysik (Vol. 1), 1860. (Translated as Elements of Psychophysics. New York: Holt, Rinehart and Winston, 1966).
17. Green, G. M. & Swets, J.A. Signal Detection Theory and Psychophysics. New York: Wiley, 1966.
18. Guilford, J.P. Psychometric Methods (2nd ed.). New York: McGraw-Hill, 1954.
19. Jennrich, R.I. & Sampson, P.F. Application of stepwise regression to nonlinear least squares estimation. Technometrics, 1968, 10, 63-72.
20. Miller, G.A. Sensitivity to changes in the intensity of white noise and its relation to masking and loudness. J. Acoust. Soc. Amer, 1947, 19, 609-619.

21. Oberlin, K.W. Variations in intensive sensitivity to lifted weights. J. Exper. Psychol., 19, 438-455.
22. Rao, C.R. Advanced Statistical Methods in Biometric Research. New York: McGraw-Hill, 1952.
23. Stevens, S.S. Psychophysics. New York: Wiley, 1975.
24. Swets, J.A. Is there a sensory threshold? Science, 1961, 134, 168-177.
25. Swets, J.A., Tanner, W. P., & Birdsall, T.G. Decision processes in perception. Psychol. Rev., 1961, 68, 301-340.
26. Teghtsoonian, R. On the exponents in Stevens' law and the constants in Ekman's law. Psychol. Rev., 1971, 78, 71-80.
27. Teghtsoonian, R. Range effects in psychophysical scaling and a revision of Stevens' law. Amer. J. Psychol., 1973, 86, 3-27.
28. Teghtsoonian, R. Psychophysics: introduction to its perceptual, neural, and social prospects (Review of S.S. Stevens Book). Amer. J. Psych, 1975, 88, (4), 677-678.
29. Thomas, D. J., Majewski, P.L., Ewing, C.L., & Gilbert, N.S. Medical qualification procedures for hazardous-duty aeromedical research. AGARD Conference Proceedings No. 231. London: AGARD, 1977, A3, 11-13.
30. Thurstone, L.L. The phi-gamma hypothesis. J. Exp. Psychol., 1928, 11, 293-305.
31. Thurstone, L.L. The Measurement of Values. Chicago: University of Chicago Press, 1959.
32. Wald, A. On a statistical problem arising in the classification of an individual into one of two groups. Ann. Math. Statistics, 1944, 15, 145-162.
33. Winer, B.J. Statistical Principles in Experimental Designs. New York: McGraw-Hill, 1962.
34. Woodrow, H. Weight discrimination with a varying standard. Amer. J. Psych., 1933, 45, 391-416.
35. Woodworth, R.S. Experimental Psychology. New York: Henry Holt, 1938.
36. Woodworth, R.S. & Schlosberg, H. Experimental Psychology. New York: Holt, 1954.
37. Wuest, F.J. Psychophysical Measurements from Two Theoretical Viewpoints. Unpublished doctoral dissertation. Brown Univ., 1961.

PROCEEDINGS OF THE HUMAN FACTORS ASSOCIATION OF CANADA MEETING, ONTARIO, CANADA, SEPT. 1980

A FACTOR ANALYTIC APPROACH TO DETERMINING STABILITY OF HUMAN PERFORMANCE

Richard H. Shannon, Ph.D.
Naval Biodynamics Laboratory, New Orleans, LA 70189

ABSTRACT

Within industry, the three measures of stability (means, standard deviations, intertrial correlations) can function as indicators of the lowering of productivity. Constant means and standard deviations can be determined using confidence intervals. Correlational equality can be concluded from the maximum deviation point between expected and observed cumulative distributions of the squared task deviations $(T)^2$. Task definition is defined as the average of the intertrial correlations of any day with all other days. The square of the loadings on each day, which were determined by a one factor solution utilizing factor analysis, were found to be similarly distributed as $(T)^2$. The recommended stability levels for future analyses are 99 percent confidence with a constant slope correction and a .650 task definition.

INTRODUCTION

Performance stability is an important concept in both the experimental and industrial environments. Presently, this construct is helping to develop a performance battery (PETER, Performance Evaluation Tests for Environmental Research), which will eventually be used to study behavior under unusual and adverse conditions. The usefulness of a test for this purpose is determined by the unchanging, stable scores in the baseline or controlled condition. This criterion is important because any effect associated with repeated measurement would be confounded with changes of performance due to the environment. Stability (Jones, 1979) is defined as the period when (1) mean performance reaches nearly constant slope over time, (2) between subject variances are homogeneous over time, and (3) relative performance standings of the subjects, reflected in cross-session reliabilities, are constant over time.

The implications of this research can be generalized to the industrial workplace. For example, the statistical properties of stability have application when learning curves are utilized as tools of management for purposes of scheduling, productivity, training and forecasting (Moore, Jablonski, 1969). With practice, people improve their ability to do work, which can be evidenced by increases in such diverse skills as scanning rate and discrimination, memory and rule-using, time-sharing and planning, movement efficiency and precision. As workers gain experience either through formal training or on-the-job exposure, their productivity increases rapidly at first; but then as performance on a particular job or task is optimized, the learning curve flattens or levels off. This flattening period is synonymous with stability. The valid determination of this property as it relates to production levels and the daily reliability of labor is critical to forecasting and scheduling within an organization.

Another tool of industry with which stability has application is that of control charts. Manufacturing processes, even when controlled, have a certain amount of variability which cannot be eliminated. When this variability is confined to random or chance variation, the process is considered to be within statistical control (Miller, Freund, 1965). This period of control is synonymous with stability. Control charts consisting of a central line and upper and lower limits for the mean, standard deviation and the range can be utilized for the purpose of detecting serious deviations from stability. These limits are determined by setting statistical confidence bands ($\pm$ 3) around the estimated population mean and standard deviation. These estimated values are usually derived by averaging the statistics of the samples collected during the period of process control. By plotting the results obtained from the samples, the determination of stability can be judged by the number of values inside or outside of the confidence limits. Although control charts usually have application to equipment variables, they are quite suited to the analyses of worker variables.

In the two industrial applications of stability just mentioned, learning curves and control charts, the emphasis is upon means and standard deviations. Intertrial correlations, however, are just as important because they give the investigator a measure of internal reliability. For example, a theoretical group of workers, who are performing a particular task, may decide to cooperate in the lowering of their production levels during baseline data collection. If the means and standard deviations were constant, the investigator would have difficulty in determining whether the data gave a valid indication of performance achievement and stability. However, rank-order positions on a daily basis (intertrial correlations) are more difficult to manipulate, especially when the people being observed are not aware of this subtle statistic. Because of the importance of reliability to performance, the purpose of this paper will be to discuss various methodologies which can be used to determine correlational stability. A cognitive experimental test, which was conducted at this laboratory, will function as the vehicle of explanation.

METHOD

The grammatical reasoning test (Baddeley, 1968) was scrutinized in order to determine whether it was suitable for inclusion in the PETER battery (Carter, Kennedy and Bittner, 1980). This test is purported to measure "higher mental processes." Twenty-three subjects took the test on 15 consecutive workdays in a standard environment. The grammatical reasoning test involves five grammatical transformations on statements about the relation between two letters: A and B. The five transformations are: (1) active versus passive sentence construction, (2) true versus false statement, (3) affirmative versus negative phrasing, (4) use of the verb "precedes" versus the verb "follows," and (5) sequential order of A versus B. There are 32 possible items, and they were arranged in a different random order on each day of the experiment. The subject responded with either a "True" or "False" depending upon the verity of each statement. For example, "True" is the appropriate response to the stimulus: A precedes B - AB. Subjects were allowed 1 minute to work on this paper-and-pencil test on each day of the experiment. The test was administered to the subjects in a group. Scores were the number of correct responses.

RESULTS AND DISCUSSION

The results indicated that the grammatical reasoning test is quite suitable for use in repeated measures experiments. The means and standard deviations appear in Table 1. The means increase linearly with practice

(slope = .3 correct responses/day) as confirmed by a repeated measures analysis of variance. The linear component of the days effect was statistically significant ($F(1,22) = 50.39$, $p < .0005$), and accounted for 90% of the variance attributable to days. There was no indication that the variance of grammatical reasoning scores changed over the 15 days ($F_{max}(15, 22) = 1.82$, non-significant at .05 level). In order to determine the days causing the significant deviations, 99 percent confidence limits were placed around the average mean and standard deviation of the 15 days, as in the construction of control charts (Miller, Freund, 1965). Each of the 15 days were treated as samples from the population. Using the t distribution for the means and the chi-square distribution for the standard deviations, the resulting central value (CV) with 99 percent upper (UL) and lower (LL) confidence limits were: (1) mean - 12.62 (CV), 15.66 (UL), 9.58 (LL), and (2) standard deviation - 5.06 (CV), 7.06 (UL), 3.17 (LL). Table 1 shows that none of the standard deviations and one mean (day 1) are outside these statistical boundaries. There is a good possibility, however, that if the experiment had continued, the means on the days after day 15 would have been outside the limitations. If a correction of .30 constant slope on the control chart for the mean had been utilized as a forecasting projection, this contingency would not occur and the investigator would still have had an estimation of stable performance.

Another condition which is necessary for stability is that of the intertrial correlations being constant over time. Table 1 depicts the task definition for each day, which is the average of the intertrial correlations of that day with all other days. In other words, task definition by day is an average of 14 correlations, and task definition for the matrix is a mean of 210 correlations. The task definition by matrix was .72. The Lawley test (Morrison, 1967) indicated that the intertrial correlations did not change appreciably after Day 4 ($^2(44) = 43.65$, non-significant at .05 level) but were not constant after Day 3 ($^2(54) = 83.29$, $p < .025$). Since day 15 was omitted from these analyses due to its relatively lower task definition, stability was noted from days 5 to 14. The usefulness of intertrial correlations can be demonstrated using the three indices of day 15. Since this day was an end point known to the subjects, there may have been a lack of concentration demonstrated by the task definition (.60). The high mean and stable standard deviation indicate, without the correlational information, that the day 15 sample was performing very well. However, only when the three indices are studied together does a more complete picture emerge.

The utility of the Lawley test in the determination of correlational stability is lowered by the following trait: non-significant results indicate that correlations among trials are equal, but a significant analysis does not mean that a differential change is present (Jones, 1979). To draw this conclusion, another alternative method is necessary. Such an approach may be factor analysis. This methodology operates to maximize the amount of variance shared commonly among the variables. When the variables are days and the cases are subjects, stability should be indicated by the loadings of the variables as well as the amount of variance explained by the first unrotated factor. This position is partially supported by Humphreys (1960) who believed that the correlational matrix containing variables of successive trials on the same task represented only one common factor. In addition, Corballis (1965) suggested a linear model as an alternative to the usual factor model of multiple solutions. The one factor solution presented in this paper is comparable to a linear model. Table 1 lists the factor loading on each day. These data indicate that 75 percent of the variance was explained by this analysis and that the average factor loading was .86. If days 5 to 14 were considered as the stable period as indicated by the Lawley test, the explained variance would increase to 85 percent, and the factor loadings would be near or greater than .90.

The Kolmogorov-Smirnov (K-S) goodness of fit test (Miller, Freund, 1965) was examined in order to determine whether it would be applicable to correlational stability analysis. The one-sample test is concerned with the amount of agreement between observed and expected cumulative distributions. For example, the test was utilized in order to determine whether task definition and factor analysis were attempting to explain similar constructs: the cumulative distribution of the explained variance on each day within the total matrix. In Table 1, the relative cumulative distributions of the squared task definitions and factor loadings were presented. The distributions are non-significant ($p < .05 = .073$), and therefore, can be considered identical. In fact, a multiple of 1.2 could be used to equate each daily task definition to its related factor loadings. This loading was determined by dividing .86 (average of factor loadings) by .72 (task definition by matrix). Since factor analytic results having a one factor solution and task definition appear to be similar constructs, the determination of correlational stability can rely mainly upon task definition. This conclusion was further supported by the results from four other mental tests (free recall, interference susceptibility, running recognition, and list differentation). In addition, these tests indicated that a .650 task definition may be an acceptable standard, since this value is comparable to 68 percent of the variance from factor analysis and to an average factor loading of approximately .82.

A one sample K-S test was conducted using the cumulative frequency distributions of squared task definitions (observed) and predicted values based upon 1.0 divided by 15. These predicted scores represented the theoretical distribution of stable and equal task definitions. The absolute maximum difference point in Table 1 was depicted to be Day 4, which was similar to the Lawley test. These results however were non-significant at the .1 level ($p < .2 = .058$; $p < .1 = .066$; $p < .05 = .073$). In other words, the K-S test indicated that correlational stability was arrived at on day 1. The difference between the Lawley and K-S, therefore, must be one mainly of test stringency. For the K-S test to have been significant and independently distributed, the level of significance would have been at the .20 level. In order to determine the stringency of the Lawley, another test was conducted based upon the distribution of days 5-14. The task definition by matrix for these nine days was .83. Using the K-S one sample test, the maximum difference was .012. In conclusion, it appears that the Lawley is very conservative and should be used with caution.

If the Kolmogorov-Smirnov test had indicated a significant departure at day 4, task definitions by day would again have to be computed using days 5 through 15. In other words, an average of ten correlations would represent the daily values while the task definition by matrix would be a mean of 110 correlations. The K-S test would again be utilized in order to determine whether the distributions of expected and observed squared task definitions were similar.

REFERENCES

Baddeley, A. D. A three minute reasoning test based on grammatical transformation. Psychonomic Science, 1968, 10, 341-342.

Carter, R. C., Kennedy, R. S., and Bittner, Jr., A. C. Grammatical reasoning: A stable performance yardstick. Unpublished manuscript, 1980.

Corballis, M. C. Practice and the simplex. Psychological Review, 1965, 5, 399-406.

Humphreys, L. G. Investigations of the simplex. Psychometrika, 1960, 4, 313-323.

Jones, M. B. Stabilization and task definition in a performance test battery. New Orleans, LA: Naval Biodynamics Laboratory, Contract No. N0023-79-M-5089, Final Report, 1979.

Miller, I. and Freund, J. Probability and Statistics for Engineers. Englewood Cliffs, N.J.: Prentice-Hall, 1965.

Moore, F. and Jablonski, R. Production Control. New York: McGraw-Hill, 1969.

Morrison, D. F. Multivariate Statistical Methods. New York: McGraw-Hill, 1967.

TABLE 1: MEANS, STANDARD DEVIATIONS, TASK DEFINITIONS, FACTOR LOADINGS, AND CUMULATIVE DISTRIBUTIONS FOR 15 DAYS

DAYS	MEANS	STD DEV	TASK DEFINITION (T)	CUM DIST $(T)^2$	FACTOR LOADING (L)	CUM DIST $(L)^2$	OBS CUM FREQ $(T)^2$	OBS CUM FREQ $(L)^2$	PRED CUM FREQ (P)	DIFF $(T)^2-P$
1	8.5	3.3	.56	.32	.68	.46	.04	.04	.07	.03
2	9.9	4.3	.58	.65	.70	.95	.08	.09	.13	.05
3	9.8	4.3	.71	1.16	.85	1.67	.15	.15	.20	.05
4	11.1	4.8	.69	1.64	.83	2.37	.21	.21	.27	.06
5	11.6	4.6	.78	2.25	.93	3.23	.29	.29	.33	.04
6	12.4	4.9	.79	2.87	.93	4.10	.37	.37	.40	.03
7	13.3	5.0	.76	3.45	.90	4.91	.44	.44	.47	.03
8	13.4	4.8	.77	4.04	.92	5.75	.51	.51	.53	.02
9	13.1	5.4	.79	4.67	.94	6.63	.59	.59	.60	.01
10	13.4	4.5	.72	5.18	.86	7.37	.66	.66	.67	.01
11	14.7	5.3	.73	5.71	.87	8.13	.73	.73	.73	.00
12	14.0	6.0	.76	6.29	.90	8.94	.80	.80	.80	.00
13	14.3	4.5	.75	6.85	.90	9.74	.87	.87	.87	.00
14	14.1	5.8	.81	7.50	.96	10.66	.96	.95	.93	.03
15	15.5	4.7	.60	7.86	.72	11.19	1.00	1.00	1.00	.00